Linked open data og geografiske autoritetsdata

Knut Anton Bøckman
Stig Svenningsen

Det Kongelige Bibliotek
HisKIS-seminar SDU 2016-09-14

Identifikation

Match af lokale registrer-inger med publicerede data:

- geonames
- open street maps
- GSTs AWS service
- Stednanve-base

KBB01-sys	621*a	621-a-underinddeling	url_OSM
000018748	København	Nyhavn	http://www.openstreetmap.org/node/894229676
000019987	København	Nørrebrogade, Nørrebros Runddel	http://www.openstreetmap.org/node/664257007
000019985	København	Nørrebrogade, Nørrebros Runddel	http://www.openstreetmap.org/node/664257007
000019976	København	Nørrebrogade	http://www.openstreetmap.org/way/310869916
000019963	København	Nørrebrogade	http://www.openstreetmap.org/way/310869916
000019950	København	Vester Voldgade	http://www.openstreetmap.org/way/149756673
000019918	København	Frederiksborggade edit	http://www.openstreetmap.org/way/77628809
000019856	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019853	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019853	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019846	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019846	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019827	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019804	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019804	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019804	København	Tornebuskegade	http://www.openstreetmap.org/way/1931867
000019698	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019683	København	Nørre Voldgade, Nørreport Station	http://www.openstreetmap.org/node/2343875251

Agenda: forbinde datasæt

- Formål – bedre dataudnyttelse
- Metode – Linked Open Data
- Værktøj – OpenRefine
- Kilder – geografiske autoritetsdata
- Udfordringer – skala, granularitet

Bibliotek

Indsamle Bevare Formidle

digitalisering
≠
formidling

Use case

Jeg ønsker at finde et historisk billede fra slagterierne i Kødbyen.

- Mon ikke Det Kongelige Bibliotek har det?
- Jeg finder Det Kongelige Biblioteks hjemmeside
- Jeg finder søgesystemet
- Jeg søger på Halmtorvet

Google historiske billeder halmtorvet

Alle Billeder Maps Videoer Mere Søgeværktøjer

Digital formidling

- Materiale: Digitalt og publiceret

- Tilgængelig for indeksering
- Kan findes via almene værktøj

- Alment anvendte beskrivelser
- Tilpas formidling til indhold

nogle eksempler

DPLA

- https://dp.la/map?

Haverford

- http://cope.haverford.edu/letterNetwork

Icelandic Saga Map

- http://sagamap.hi.is/is/

NYPL

- http://pic.nypl.org

Biblioteksdata transformeres til

MERE ANVENDELIGE DATA

Tim Berners-Lee: Linked data (2006)

1. Use URIs as names for **things**
2. Use **HTTP URIs** so that people can look up those names.
3. When someone looks up a URI, **provide useful information**, using the **standards** (RDF*, SPARQL)
4. Include **links** to other URIs, so that they can **discover more things.**

Linkning - http

Web of Documents (html)

- Links mellem dokumenter
- Information i tekst-strenge
- Semantik ligger i kontekst
- Lavet for mennesker; ikke maskinevenlig

Web of Data (RDF)

- Entydig reference for ting
- Entydig reference for relation
- Brug af anerkendte internet-standarder

Data

- **Resource Description Framework (RDF)**
- **Data beskrives i prædikative udsagn** (triples)
  - Subjekt **(Moesgård museum)**
  - relation **(ligger i)**
  - Objekt **(Århus)**
- Selvstændige, meningsfulde **enheder**
  - **Uafhængig** af dokumentkontekst
  - Forøget betydning ved **linkning** ☺

Strings to things

- Erstat tekstbaserede navne (subjekt +objekt) med URI'er
- Også URI for prædikatet (relationen mellem dem)

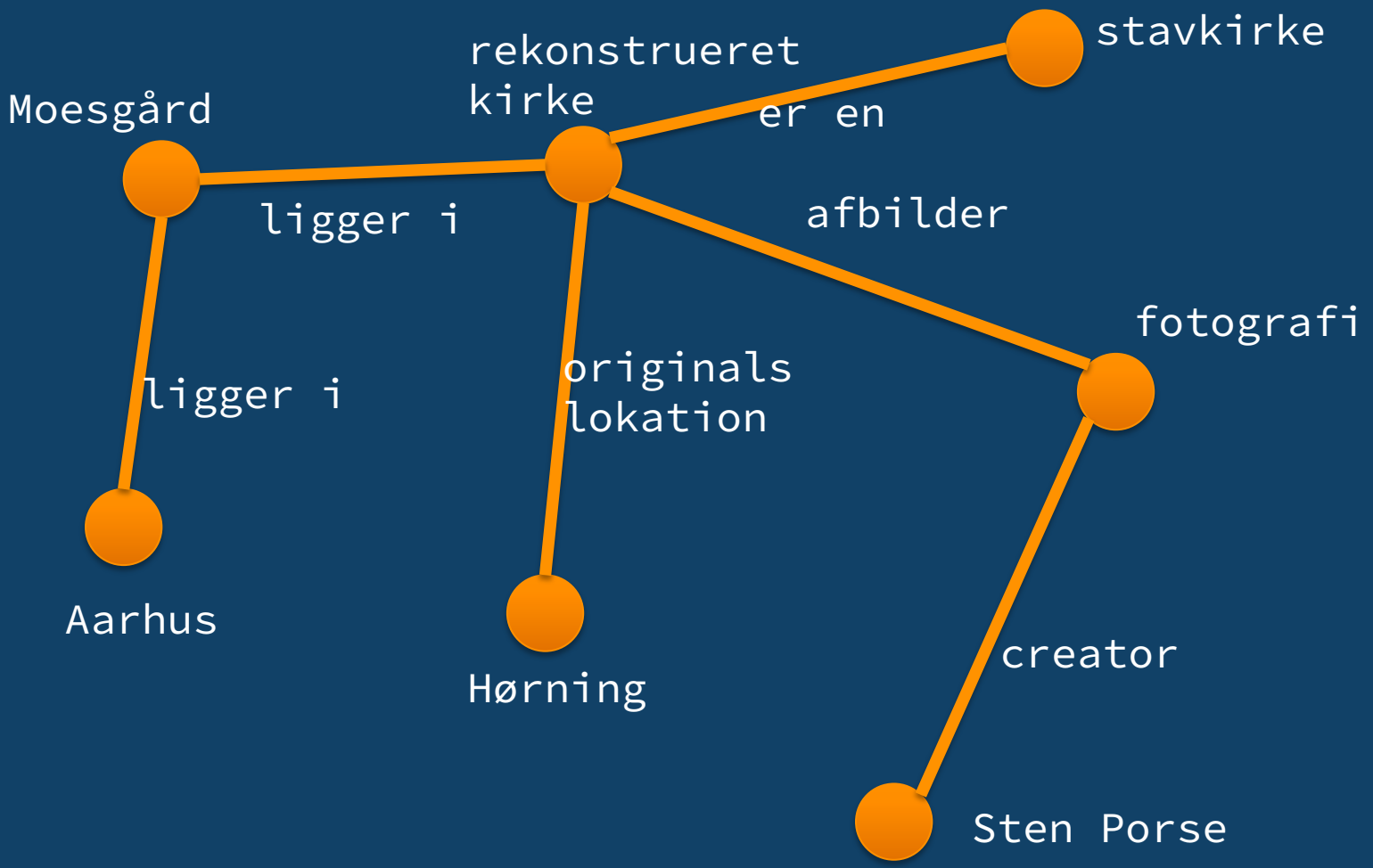
Moesgård Museum er placeret i Aarhus

<https://www.wikidata.org/wiki/Q3299384>

<http://purl.org./dc/terms/Location>

<http://vocab.getty.edu/tgn/7003442>

Linkning (tænkt eks.)



Bibliotekets beskrivelse

- Dokument der beskriver en ressource
- Format: Danmarc2
- Kodede indførsler
- Semantisk scope er dokumentet
- Data som tekststrenge

		l	skan Türck 50344
248	00	g	DT103789
		a	Kølehus. Interiør, svin
		e	foto. Sven Türck
		l	neg. Türck 43223
248	00	g	DT103790
		a	Kølehus. Interiør
		e	foto. 152470
300	00	n	Fotografi
599	00	b	KBB980603
621	00	a	København, Halmtorvet
		b	Danmark
700	00	a	Türck
		h	Sven
		f	fotograf
		c	ca. 1897-1954
700	00	a	Nielsen

Biblioteksdata (billeder)

- Strukturerede data
- Indeholder stedsbeskrivelser

- Mangler linkning til andre data

- Ikke alment udbredt format
- Ikke entydige navne på ting

Ting - navn: URI

- **Eksport af kildedata**
  - To datasæt: København og udenfor
  - csv
- **Normalisering/tilpasning**
  - OpenRefine
- **Match mod autoritetskilder**
  - Geonames, OpenStreetMaps, andre
- **Genimport af berigede data**

Håndtering af data

Eksport: csv
Load i OpenRefine

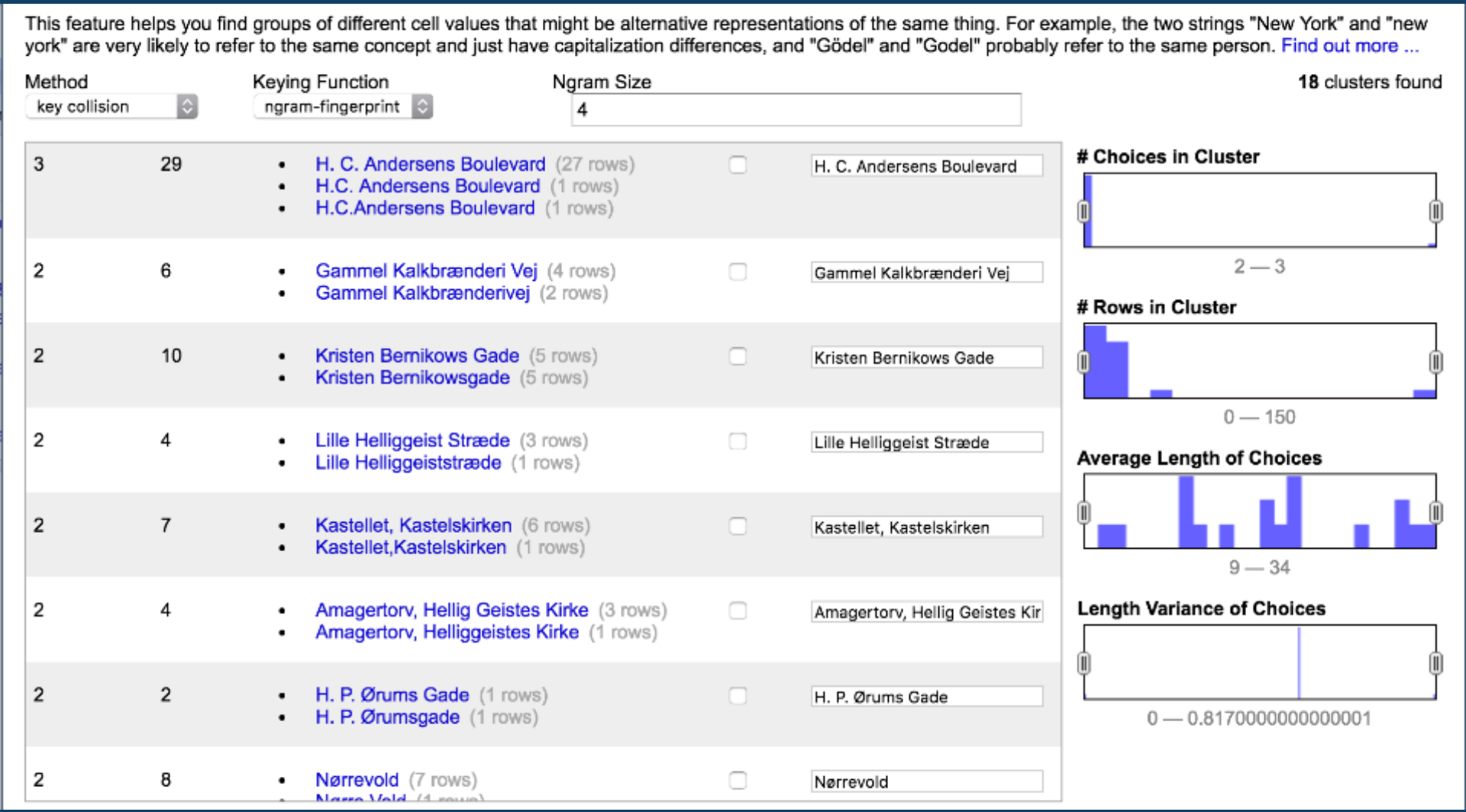
Et google regneark med (meget) ekstra funktionalitet

Grafisk UI: lav tærskel

Vedligeholdes af aktiv community

000001035			aSlangerup	bDanmark
000001350			aMünchen	bTyskland
000002435			aHälsingborg	bSverige
000002635			aSchloss Husum	bTyskland
000002735			aHobro	bDanmark
000002835			aFrydenlund Slot	bDanmark
000003935			aSønderborg	bDanmark
000003935			aTinglev	bDanmark
000004735			aKarise, Karise kirke	bDanmark
000006235				bDanmark
000006350			aWashington	bUSA

Normalisering af data



Normalisering af data

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision Keying Function: metaphone3 **191** clusters found

- Frederiksholms Kanal, Fæstningsmaterialgården (1 rows)
- Frederiksholms Kanal, Hestegardens Kaserne (1 rows)
- Frederiksholms Kanal, Materialgården (1 rows)
- Frederiksholms Kanal, Nationalmuseets Klunkehjem (1 rows)
- Frederiksholms Kanal, Wedells Palæ (1 rows)

16	207	• Kongens Nytorv (64 rows) • Kongens Nytorv, Det kongelige Teater (47 rows) • Kongens Nytorv, Charlottenborg (31 rows) • Kongens Nytorv, Hotel d'Angleterre (13 rows) • Kongens Nytorv, Hovedvagten (10 rows) • Kongens Nytorv, Magasin du Nord (9 rows) • Kongens Nytorv, Krinsen (8 rows) • Kongens Nytorv, Gjethuset (5 rows) • Kongens Nytorv, Thotts Palæ (5 rows) • Kongens Nytorv, Harsdorffs Palæ (4 rows) • Kongens Nytorv, Kanneworffs Hus (3 rows) • Kongens Nytorv, Stephan à Porta (3 rows) • Kongens Nytorv, Store Nordiske Telegrafselskab (2 rows) • Kongens Nytorv, Charlottenborg, Udstillingsbygningen (1 rows) • Kongens Nytorv, Den Militær Højskole (1 rows) • Kongens Nytorv, Å Porta (1 rows)	☐	Kongens Nytorv
11	346	• Vesterbrogade, Tivoli (177 rows) • Vesterbrogade (151 rows) • Vesterbrogade, Industriforeningen (9 rows)	☐	Vesterbrogade, Tivoli

 # Choices in Cluster

2 — 60

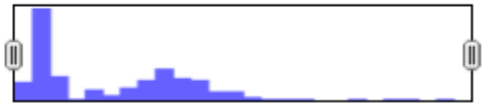
 # Rows in Cluster

0 — 350

Average Length of Choices

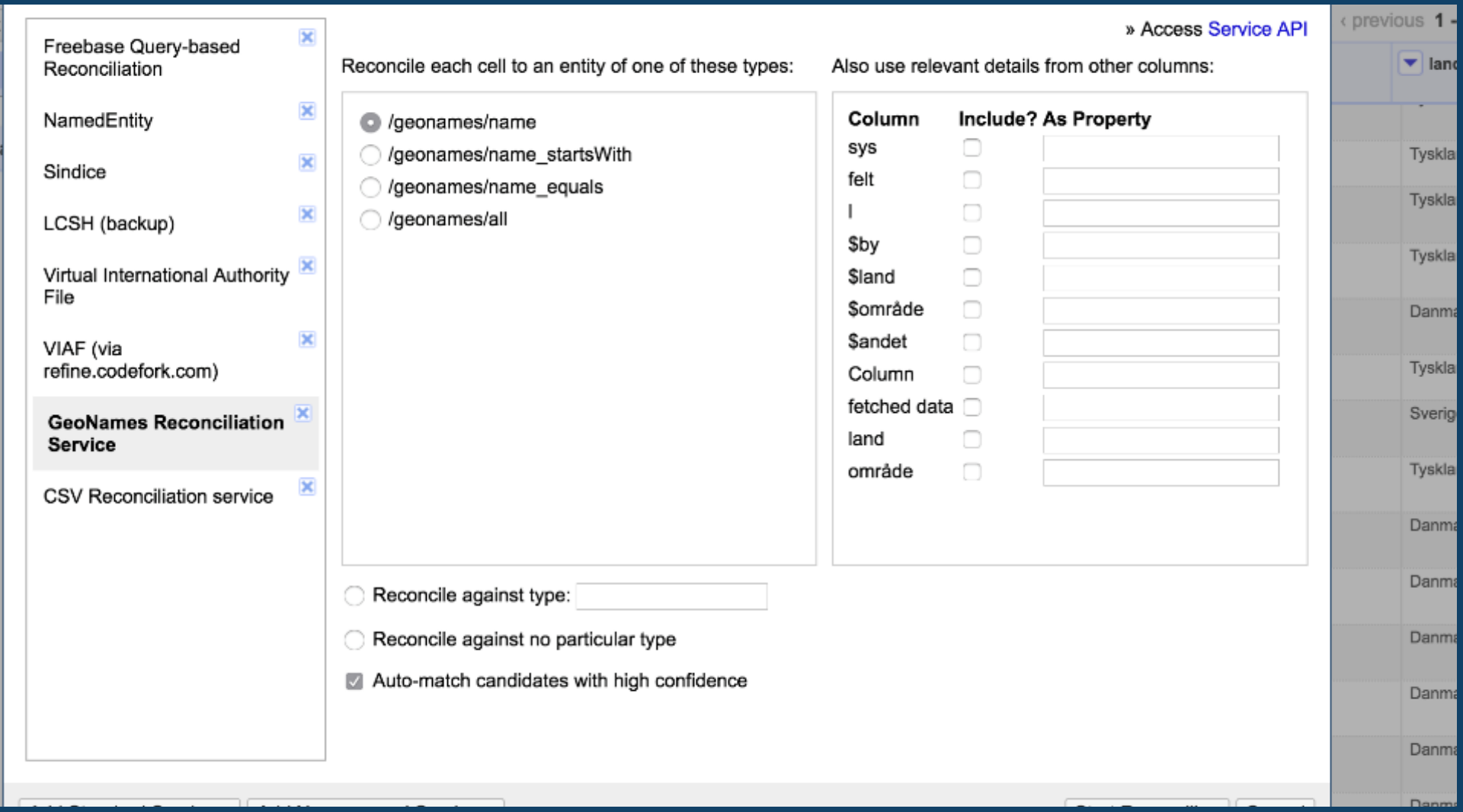
8 — 46

Length Variance of Choices



0 — 25

Match data mod andre kilder



Hent data ind

GeoNames

000000383	Greifswald | 54.09311, 13.38786 Choose new match	
000000384	Bohus | 57.85, 12.01667 Choose new match	
000000878	Varberg | 57.10557, 12.25078 Choose new match	
000001033	Slagelse | 55.40276, 11.35459 Choose new match	
000001034	Slagelse | 55.40276, 11.35459 Choose new match	
000001119	Holstebro | 56.36009, 8.61607 Choose new match	
000001176	Odense | 55.39594, 10.38831 Choose new match	
000001177	Odense | 55.39594, 10.38831 Choose new match	
000001195	Bulbjerg | 57.1575, 9.02588 Choose new match	Bulbjerg | 57.1575, 9.02588 | http://sws.geonames.org/2623248 | Danmark
000001208	Frederiksborg Slot | 55.93451, 12.30041 Choose new match	Frederiksborg Slot | 55.93451, 12.30041 | http://sws.geonames.org/2621938 | Danmark

OpenStreetMaps

KBB01-sys	621*a	621-a-underinddeling	url_OSM
000018748	København	Nyhavn	http://www.openstreetmap.org/node/894229676
000019987	København	Nørrebrogade, Nørrebros Runddel	http://www.openstreetmap.org/node/664257007
000019985	København	Nørrebrogade, Nørrebros Runddel	http://www.openstreetmap.org/node/664257007
000019976	København	Nørrebrogade	http://www.openstreetmap.org/way/310869916
000019963	København	Nørrebrogade	http://www.openstreetmap.org/way/310869916
000019950	København	Vester Voldgade	http://www.openstreetmap.org/way/149756673
000019918	København	Frederiksborggade edit	http://www.openstreetmap.org/way/77628809
000019856	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019853	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019853	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019846	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019846	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019827	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019804	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019804	København	Fiolstræde	http://www.openstreetmap.org/way/12645065
000019804	København	Tornebuskegade	http://www.openstreetmap.org/way/1931867
000019698	København	Nørre Voldgade	http://www.openstreetmap.org/way/225625919
000019683	København	Nørre Voldgade, Nørreport Station	http://www.openstreetmap.org/node/2343875251

Kilder til geografiske

AUTORITETSDATA

Kilder (1/2): LOD

Geonames

- Åben API
- Nem rekoncilering
- Åbne data
- Linked data
- Stabil URI
- Dårlig detaljegrad for gadenavne i København

Open Street Maps

- Begrænset API
- Åbne data
- Leverer data i JSON, skal parses
- Ingen URI-policy
- God detaljegrad for gadenavne i København
- Variabelt datagrundlag

Kilder (2/2)

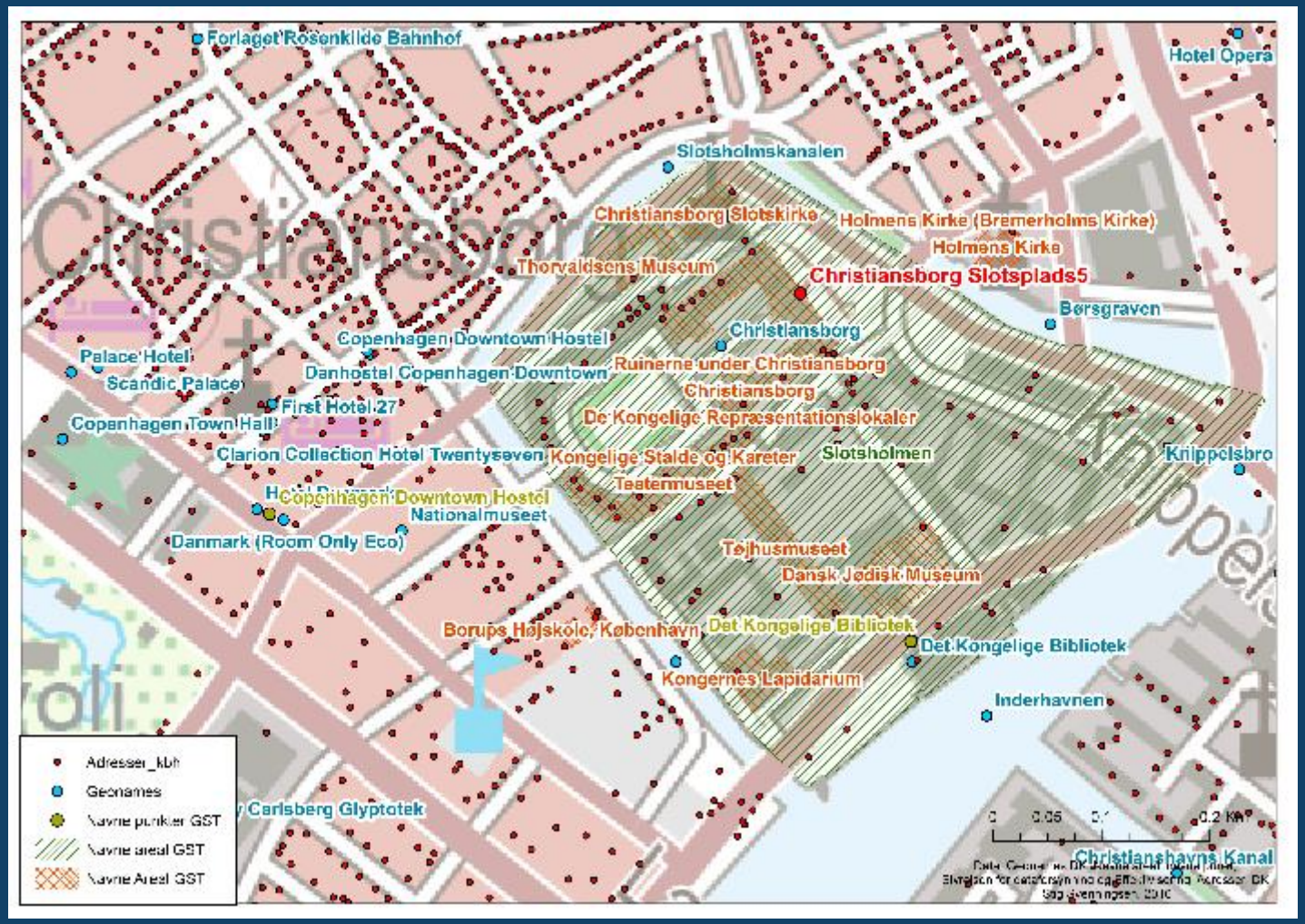
GST adresse-udtræk fra aws.dk

- Ikke linked data
- Ingen åben API
- Data åbne?
- Datadump med identificerede indførsler på husnummer-niveau

Stednavnebase fra Afdleling for navneforskning

- Historiske stednavne
- Meget detaljeret udenfor København
- Ikke detaljeret i København
- Ikke publiceret
- Ikke åbne data
- Ikke linked data

Udfordringer med LOD - fra en geografs perspektiv



Udfordringer med brugen af geografisk sted data

- Geometri for den geografiske reference
  - Punkt.
  - Linje.
  - Polygon.
- Skala
  - Christiansborg: område, stednavn, begreb, adresse(r)
- Præcision
  - Det Kongelige Bibliotek i to punkter.
- Bias
  - Geonames mest butikker.

referencer

- Tim Berners-Lee (2006): Linked data https://www.w3.org/DesignIssues/LinkedData.html
- OpenRefine: http://openrefine.org
- Blog om LOD i DEFF-projektet Linked, Open & Social: http://los3blog.wordpress.com/
- Se også eksemplerne i slide 9.